PepsiCo and La Trobe University release annotated gene set and associated files for OT3098 v2 genome in partnership with GrainGenes

By: Dr. Mandy Waters, PepsiCo & A/Prof Mathew G Lewsey, La Trobe Institute for Agriculture and Food

Description of data release

PepsiCo & the La Trobe Institute for Agriculture and Food are releasing a set of predicted transcripts mapped to OT3098 v2 reference and their associated annotations. Researchers can access the annotation file as a genome browser track. Additional files including a GFF, transcript nucleotide fasta, and an Excel file with associated gene annotations are also available *via* the download site on GrainGenes.

List of files:

PepsiCo_OT3098_V2_panoat_nomenclature.gff3: gff file with predicted genes (details below on prediction pipeline and nomenclature)
PepsiCo_OT3098_V2_panoat_nomenclature_cDNA.fasta: fasta file with nucleotide sequences for predicted transcripts
PepsiCo_OT3098_V2_panoat_nomenclature.xlsx: annotation file in Excel format.

Use:

Researchers are free to use and publish with all OT3098 genomic resources shared on GrainGenes. A detailed publication is forthcoming but researchers can freely use these resources as they become available:

- Genome Browser: https://wheat.pw.usda.gov/jb?data=/ggds/oat-ot3098v2-pepsico
- Data Download: https://wheat.pw.usda.gov/GG3/graingenes-downloads/pepsico-oat-ot3098-v2-files-2022

Citation:

If you use these resources, please cite the following until a detailed formal publication is made:
"*Avena sativa* – OT3098 v2, PepsiCo, https://wheat.pw.usda.gov/jb?data=/ggds/oat-ot3098v2-pepsico"

Pipeline information:

Written by Dr. Changyu Yi, Bioinformatics Analyst, La Trobe University

- **Obtaining high-quality transcripts**

Two Iso-seq runs were performed. The first dataset was generated from six tissues, including root, young leaf, meristem, germinated seeding, whole oat and developing seed (20 days after anthesis). The second dataset was generated from shoot and root tissues harvested 22 days after sowing. The raw sequencing data was processed using the Isoseq 3 package to obtained high-quality (HQ) transcripts. The HQ transcripts from both Iso-seq datasets were mapped to the oat reference genome (OT3098 v2) using pbmm2.

- **Removing redundant transcripts**

The software Transcriptome Annotation by Modular Algorithms (TAMA) was used to remove redundant transcripts and merge the two Iso-seq datasets. Transcripts with mapping coverage and identity lower than 95% were firstly discarded, and the remaining transcripts were then collapsed and merged if the difference in 5 prime, 3 prime and exons of these transcripts were less than 500bp, 100bp and 20bp, respectively.

- **ORF prediction and gene annotation**

The coding sequence of each transcript was predicted using TransDecoder. Subsequently, the transcript sequence and their TransDecoder predicted protein sequence were annotated using Trinotate software with Swiss-Prot, Pfam and UniRef90 databases. In addition, the predicted protein sequences were annotated against the evolutionary genealogy of genes: Non-supervised Orthologous Groups (eggNOG) database using eggNOG-mapper. We also annotated potential transcription factors using the PlantTFDB database.

Most genes have more than one transcript and it is possible that different transcripts of the same gene might be annotated to different entries in the queried databases. To generate a gene-level annotation, a representative transcript for each gene was selected, and the annotation of the representative transcript was transferred to the corresponding gene. To this end, we favoured the Swiss-Prot and Pfam annotation results. Transcripts that were annotated with most common Swiss-Prot ID were retained for next step. Next, the transcripts without Pfam annotation were discarded, and lastly, the transcript with lowest Blast E value against Swiss-Prot were selected as the representative transcript.